基于条件随机场的汉语词汇特征研究 *

黄定琦，史晟辉

(北京化工大学 信息科学与技术学院，北京 100029)

摘 要：汉语语言在书面表达时不具有天然分词的特性，词汇与词汇之间没有分词标记，因此在汉语文本的识别中需结合其行文的习惯及规则，即所谓的词汇特征。已有研究通常在实验中显式地标注词汇特征来提高识别效果，增加了人工处理流程，极大地加重了算法移植的工作量。研究并归纳了常用汉语语言的词汇特征，并利用条件随机场（conditional random fields，CRF）的特征提取能力，自行实现了复杂特征函数，在语料只具有简单标注的前提下，隐式地提取词汇特征，提高了识别效果。实验证明，在汉语分词中应用复杂词汇特征能有效提高识别性能，提供了在应用中提高识别算法可移植性的新思路。

关键词：条件随机场；汉语；词汇特征；信息提取

中图分类号：TP391 doi: 10.19734/j.issn.1001-3695.2018.10.0859

Study of Chinese lexical features base on conditional random fields

Huang Dingqi, Shi Shenghui

(*College of Information Science & Technology*, *Beijing University of Chemical Technology*, *Beijing* 100029, *China*)

Abstract: In Chinese written expression, there is no word segmentation between vocabularies, so the principle of writing (or called lexical features) is what it needs to process the segmentation of Chinese content. Former researches usually mark the lexical features into training content to improve the performance, which increases the manual processing flow and the workload of the algorithm transplantation. Based on Conditional Random Fields (CRF) and the simple tags, this paper improves the recognition performance by concluding the lexical features of Chinese and transforming them to complicated functions which used by CRF. Experiments show that applying complex lexical features in Chinese word segmentation can effectively improve recognition performance and provide a new way to improve the portability of recognition algorithms in applications.

Key words: conditional random fields; Chinese; featuring Chinese word; information extraction

0 引言

随着机器学习的发展，信息提取（information extraction，IE）技术已经逐渐被应用在各个领域的生产过程中，建立各个领域的智能知识库成为越来越迫切的需求。而其中从各式各样的文字说明、描述中提取信息关键词，即命名实体识别（named entity recognition，NER），是一项必备而重要的工作。

关于汉语命名实体识别的研究，目前主要集中于特定领域中的特定格式文本的特定关键字识别。例如，文献[1]基于医学领域中的中文临床病历内容，研究病历记录中的疾病、症状以及时间记录等关键信息；文献[2]从模型方法对比、特征标注集的角度，研究从医药领域的药品说明中抽取症状信息的自动化方法，发现增加有效标注信息能够提升信息抽取的准确率及查全率；文献[3]从使用模型加自定义规则的方式，识别文本中的时间表达式；文献[4]则研究在电子商务领域中的产品描述信息的自动分词。

以上研究着重考虑特定领域的指定信息特征，其提取方法对标注集的依赖性较强，在其他领域应用该方法时，均需要花费大量的精力构建标注集。因此，本文研究了基于条件随机场（conditional random fields，CRF）的汉语语言特征，在汉语文本只具有简单标注的条件下，尝试从中提取其行文特征，寻找汉语文本的词汇特征，使其能够在简单标注的语料下都能实现较好的性能。

1 相关研究概述

1.1 条件随机场（conditional random fields，CRF）

CRF[5]是一种统计模型。由隐马尔可夫模型（hidden Markov model，HMM）理论与最大熵模型（maximum entropy，ME）理论发展而来，同时具有该两种模型的优势。相比于 HMM，CRF 可以使用复杂的特征函数对指定信息进行评判，在训练过程中，能够充分利用特征函数指定的文本上下文信息进行参数调整，从而在推理分词过程中应用上下文信息。

基于这个优势，CRF 大量被应用于各种场景中。文献[6]在影像分类中应用了 CRF；文献[7]将 CRF 应用于英文法律文档（legal documents）中，识别文档中的标语以及找出文档的判决先例（precedence）；文献[8]研究了如何在混合语言的文本中精确地识别每个单词的所属语言，使用 CRF 来辅助算法提高识别效果；文献[9]甚至将 CRF 应用到了社交网络的社区团体的识别中。

CRF 的核心原理为

$$p(y|x_i)\propto \exp\left(\sum_{k}^{K}\omega_k f_k(y,x_i,\boldsymbol{X})\right)$$

其中：$\boldsymbol{X}$ 表示当前观测序列；$x_i \in X$ 表示当前观测值；$\boldsymbol{Y}$ 表示

收稿日期：2018-10-31；**修回日期：**2019-01-15 **基金项目：**北京市教委项目（GWGJ201608）

作者简介：黄定琦（1992-），男，广西百色人，硕士研究生，主要研究方向为文本信息提取、文本实体识别（knypys@foxmail.com）；史晟辉（1974-），女，副教授，博士，主要研究方向为自然语言处理、编译技术、医学数据的标准化、生物信息.

所有标记集合；$y\in Y$表示标记值；K表示特征函数的总个数，下标 $k<K$ 表示特征函数 f 与相应权值 ω 的对应元素下标，而 K 则表示模型所包含的特征函数总数。从上式可以看出，在 CRF 的应用中，对于指定观测序列 $\boldsymbol{X}$，将某个观测值 x_i 标记为 y 的概率与模型包含的所有 K 个特征函数 f 及相应权值 ω 成正比关系，即特征函数的选取对模型应用效果有着决定性的作用。

而对于有关 CRF 的研究，文献[10,11]对如何进行特征函数的选取做了对比研究，证明了在 CRF 的应用场景中，选择适合的特征函数能够明显提高识别效果。因此，在中文命名实体识别的过程中，如何选取相应的语言特征函数，是将 CRF 应用于该过程中的一个重要问题。CRF++是现有的在研究中常用的模型工具之一。

1.2 汉语语言特征

依存语法（dependency grammar，DG）[12]又称从属关系语法，是一种建立在词汇之间的关系上的，结合了语言语义与语言表达的语法理论，为语言语义研究提供一种思路。这种结构思想已经被许多研究者接纳，如文献[13]利用 DG 规则识别汉语里的复句关系；文献[14]探索了如何构建 DG 结构的汉语树库；文献[15]在词纠错中应用了依存语法。

基于 DG 理论，句子中的词语共同构成了该语句的中心思想，并在其中起不同的支撑作用。因此，语言特征可以看做特定的表达框架，语言识别过程正是梳理整句话包含的所有词汇的框架的过程。此外，在汉语语言中还存在一个字成词的过程。有些字能够单独表达意思，而有些字却要组合起来才有实际意义。因此在对汉语语料进行处理的过程中，还要综合考虑词汇特征，结合汉语成词方式，共同对句子的框架进行梳理。

属于不同领域的描述文本均有其特殊的表达方式与行文结构，但这些特殊的差异并不能有悖于原始语言的约束。不同的汉语文章，虽然其表达方式、核心思想的差异巨大，但其表达架构将毫无疑问采用汉语语言约定俗成的规范，因此其在语言层面上均具有相同或相似的表达特征。

1.3 算法移植性问题

在使用机器学习方法对文本进行识别的过程中，通常都是一个文本标记的过程。将文本中的文字分别进行识别标记，最终再按标记的定义转换成分词的信息，从而达到提取文本信息的目的。

在多数的研究中，为了提高识别的准确率，经常会采用加入额外标记[16]的方法辅助模型进行内容识别。比如，在语料中加入词性的标记（标注），使得识别时多了一项可参考特征，从而达到提高识别效果的目的。因此自动标注算法[17,18]也成为了研究人员的方向之一。即便如此，标记的增多还是固化了模型的性能方向，因此其最大的问题在于训练语料库的构建，模型在运行中会极大地依赖训练语料中的信息，导致其在不同领域文本中识别的时候不能得到快速的运用，需要依赖指定的标注，降低了算法的可移植性。

为了解决识别算法的移植性问题，在研究中，本文将不引用任何附加标记，仅将词汇分为单字词与多字词两类词汇，将标点符号也归类为单字词，尽可能减少实验需要参考的额外信息，进行词汇特征的提取实验。

2 基于条件随机场的汉语词汇特征

结合 DG 理论，汉语词汇可以分为两大类：a)在句子中影响框架结构的词语，如动词、介词、否定词，这些词语形成了句子中心思想的框架，决定了文中描述实体之间的关系；(b)处于框架基础的词语，如名词、代词，这些词指明了句子的描述实体。

在研究词汇之间的关系之前，需要先将字组成词，而不同类型词汇的出现频率与特征都是不同的。影响句子意义结构的词往往在表述中经常出现，属于汉语中的常用关键词，所以其词汇特征也较明显；相反，第二类词在不同领域的语料中出现的频率往往不同，大多只有具备了相应知识才能识别的词汇，因此这类词的词汇特征并不明显。

本文主要研究内容为归纳不同词汇特征对识别结果的影响，并从中找出最优特征集，保证识别结果的有效性。

2.1 基于条件随机场的特征研究

在条件随机场中，其特征函数以及函数相对应的权重起决定性作用。其中，特征函数则是模型的工作中心，条件随机场通过用户选择的特征函数，对训练语料进行特征提取，即权重调整。

特征函数是特征的具体表现形式，本文对一些在表达中常出现的汉语词汇特征进行归纳总结，并将其转换为能够为条件随机场所用的特征函数形式。在相同的训练语料环境下，研究其在识别上的效果，从而找出对于汉语词汇识别较为敏感的特征集。

2.1.1 特征函数

在条件随机场的理论研究中，通常使用字母 $\boldsymbol{f}$ 表示特征函数，字母 $\boldsymbol{\omega}$ 表示特征函数的权值。其中 $\boldsymbol{f}$ 由用户选取的特征函数提取模板 $\boldsymbol{F}$ 生成；$\boldsymbol{f}$ 与 $\boldsymbol{\omega}$ 与所用的训练集均有关，且

$$\omega \propto f(y, x_i, \boldsymbol{X}) \cdot \tilde{p}(y, \boldsymbol{X}_f)$$

$$\tilde{p}(y, \boldsymbol{X}_f) = \frac{\text{count}(y, \boldsymbol{X}_f)}{\text{count}(\boldsymbol{X}_f)}$$

其中：$\boldsymbol{X}$ 表示观测序列；x_i 表示当前观测值；y 表示标记值；$\boldsymbol{X}_f$ 则表示与特征 f 的描述直接相符的序列段；函数 count 则是次数函数，若当前观测值 x_i 与周围指定的其他观测值 $x_j\in\boldsymbol{X}$ 符合特征 f 的描述，则函数特征 f 返回 1，否则返回 0 表示当前观测位置 i 不具有特征 f。也就是说，若有标记值为 y_l 的观测值 x_l 使得 $\boldsymbol{X}_f\neq\varnothing$，则

$$f(y_l, x_l, \boldsymbol{X}) \equiv 1$$

在模型的训练中，模型会逐个匹配训练集中的观测序列，对符合特征 f 的序列段进行记录，并调整对应 ω 的值，使得 ω 大体上与该特征出现的概率成正比。通常在 CRF 的应用中，先会指定一个特征提取模板，再结合训练集的分布生成特征函数。

2.1.2 特征函数提取模板

特征函数 f 一般以抽象数据的形式存在于模型中，而特征函数提取模板 F 则是由用户根据功能需要，按需求指定的一些特征提取规则。即有

$$F(x_i, \boldsymbol{X}) \xrightarrow{L} \begin{pmatrix} f_0(y, x_i, \boldsymbol{X}) \\ f_1(y, x_i, \boldsymbol{X}) \\ \vdots \\ f_K(y, x_i, \boldsymbol{X}) \end{pmatrix}$$

其中：$\boldsymbol{X}_L=\{\boldsymbol{X}_0, \boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_l\}$表示训练集中的各个观测序列集合，指定好模板 F 的提取规则后，通过遍历训练集中的序列，能够得到 K 个互不相同的特征函数 f。通常，模型中会同时使用到多个特征函数提取模板。

如图 1 所示，特征函数提取模板表示的是当前标记 y_i 与指定观测值集合 $\boldsymbol{X}_t\subseteq\boldsymbol{X}$ 的某种规则关系，图中为 $\boldsymbol{X}_t=\{x_{i+2}\}$。在实际应用当中，模板 F 记录了标记值与观测值集合的相对位置关系，而当前标记 y 与指定观测值集合 $\boldsymbol{X}_t$ 均为已知固定

值，因此，在不同的训练序列 $\boldsymbol{X}$ 中就可以形成许多具体的特征函数 f。在此，特征函数描述了从训练集中，以指定模板规则从训练集中提取出来的具体观测序列。

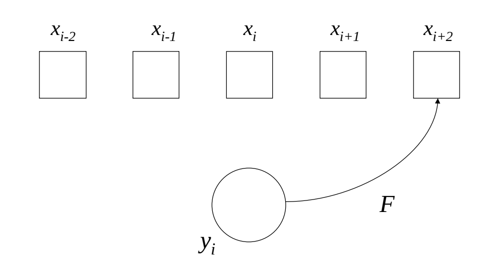


图 1　特征函数提取模板示意图

Fig. 1　Template of feature extracting function

2.2　基于条件随机场的词汇特征

条件随机场中的特征函数 $f(y, x_i, \boldsymbol{X})$ 是一个描述标记 y 与序列 $\boldsymbol{X}$ 之间的关系的函数，因此要使用条件随机场，也需要将特征定义为相同的描述形式，即定义特征函数提取模板 $F(x_i, \boldsymbol{X})$。本节将解释具体的词汇特征，对常用的汉语词汇的特征进行归纳总结，并将其转换成为描述标记与序列之间关系的特征函数。

2.2.1 随机特征（random features，R）

随机特征即随机指定的特征，指的是当前观测值与指定的另一个观测值的关系。其使用固定相对位置选取单个观测值作为目标观测集合 $\boldsymbol{X_t}$，即 x_j 的值。通常令 $j=i+n$，$n \in \mathbf{N}$，其中 $\mathbf{N}$ 为整数集。即取临近的观测值作为目标观测集合，若目标值 x_j 不存在，则将观测值指定为特殊标记 OUT，如：

$$\boldsymbol{X}_t = \begin{cases} \{x_j\}, x_j \in \boldsymbol{X} \\ \{\mathrm{OUT}\}, x_j \notin \boldsymbol{X} \end{cases}$$

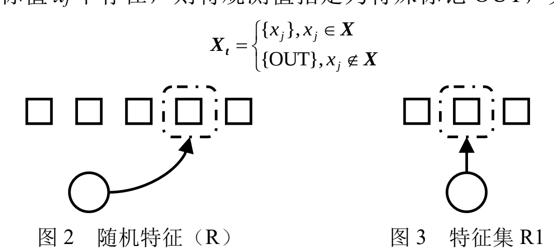


图 2　随机特征（R）　　图 3　特征集 R1

Fig. 2　Random feature（R）　　Fig. 3　feature set: R1

如图 2 所示，图中的箭头表示一个特征，每个方形表示当前观测序列 X 中的字 x_i，圆形表示当前标记 y，即当前标记与箭头指向的观测值（序列）存在特征关系。对于“基于条件随机场的词汇特征”这一表述中，其中的“特”字就具有前一字是“汇”，后一字是“征”这样的特征，显然其观测字与当前字的相对位置是由用户自行指定的。而之所以称之为随机特征，是因为在实际应用中，当前标记与指定未知的观测字并非一定存在特征关系，从而影响模型对特征的提取。

本文将用字母 R 加数字的组合表示常用的随机特征集合，其中数字表示对称窗口的宽度，R1 如图 3 描述，R5 如图 4 描述。显然，R1 中只有一个随机特征，而 R5 则是具有 5 个随机特征的集合。

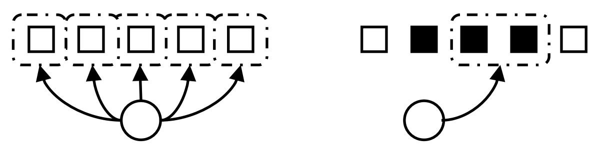


图 4　特征集 R5　　图 5　后缀特征（P）

Fig. 4　Feature set: R5　　Fig. 5　Postfix feature（P）

2.2.2 词缀特征（prefix & postfix feature，Pre & Post）

词缀特征即词汇的组成特征，指的是在处于词汇中的单字与其他字的关系。若以 $\boldsymbol{x_m}$ 表示当前观测值 $\boldsymbol{x_i}$ 所属词汇的词边界上的观测值，则有

$$\boldsymbol{X}_t = \begin{cases} (x_i, x_m], \mathrm{Postfix} \\ [x_m, x_i), \mathrm{Prefix} \end{cases}$$

即词缀特征的目标观测集合 $\boldsymbol{X_t}$ 由当前观测值 $\boldsymbol{x_i}$（不包含）与词边界 $\boldsymbol{x_m}$（包含）之间的观测值构成，分为前缀与后缀两种不同的形式。为了明确词缀特征的信息，在应用中需要判断序列是否为词边界，因此要求训练集选取的标记集中包含足够可以判断词边界的信息。

同样在“基于条件随机场的词汇特征”这个表述中，根据汉语语言特性，就可分为“基于”“条件随机场”“的”“词汇”“特征”五个具有实际表述意义的词汇，而“场”字就具有“条件随机”这样的前缀特征，“随”则同时具有前缀“条件”以及后缀“机场”。

如图 5 所示，箭头依然表示一个特征，方形仍表示当前观测序列（部分），而相邻实心方形表示三个观测值能够组成有实际意义的词汇，根据词缀特征的描述，显然仅有词汇中的字有词缀特征。特殊地，常用的词缀特征有前缀（pre）特征和后缀（post）特征，图 5 则描述了一个三字词的第一字的词后缀特征。

2.2.3 邻接特征（antecedent & subsequent feature，Ant & Sub）

邻接特征表示与当前表达相邻的实体，即词边界外的词。若以 $\boldsymbol{X_n}$ 表示与当前观测值 x_i 所属词汇邻接的词汇（包括单字词与标点符号），则有

$$\boldsymbol{X}_t = \boldsymbol{X}_n$$

在“基于条件随机场的词汇特征”这个表述中，“随”字包含在词语“条件随机场”内，具有与先行词“基于”、后继词“的”相邻的邻接特征。基于 DG 理论，词与词之间，尤其是相邻词之间存在着特殊的关系，而这些关系能够给分词边界提供特征信息。由于词汇有两种不同的邻接状态，即邻接特征分为两种，先行词（Ant）特征与后继词（Sub）特征，如图 6、7 所示。

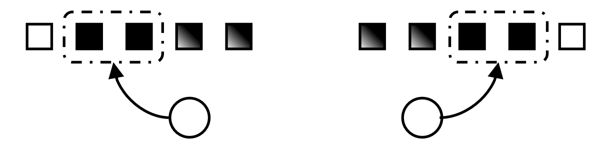


图 6　先行词(Ant)特征　　图 7　后继词(Sub)特征

Fig. 6　Antecedent feature　　Fig. 7　subsequent feature

图 6、7 中，相同填充的方形表示其处于同一词汇中。顾名思义，仅有词汇的边界字有邻接特征。

2.2.4 边界特征（marginal feature，MS & ME）

词汇的边界特征，表示词汇的开始字或结束字。边界特征有两类，一是开始特征（MS），二是结束特征（ME），描述当前字作为词边界的概率大小。例如“条件随机场”中的“条”具有开始字特征，而“场”具有结束字特征。

以 x_m 表示当前观测值 x_i 所属词汇的词边界上的观测值，则有

$$\boldsymbol{X}_t = \{x_m\}$$

边界特征在识别中通常只起辅助作用，为长词汇的边界提供判断依据。边界特征与邻接特征相似，仅有词汇的边界字具有边界特征，但边界特征描述的是当前词的结束（开始）特征，而邻接特征主要考察与当前词相邻的词汇。

2.2.5 构词特征（dictionary features，D）

构词特征描述的是不同词汇之间的共同之处，即构造词汇的规则。构词特征主要用于在识别中辅助判断指定文字序列是否具备相应的构词规则，将训练语料中的规则运用到识

别中。再者，构词特征与基于词典的识别有异曲同工之处，而构词特征比词典更抽象，将词典与规则相结合，为词汇的识别提供依据。

令 $D_m=\{d_0, d_1, d_2, \cdots, d_m\}$, m∈N 表示某种构词规则序列，函数 $g(x, d)$为二元函数，当且仅当 x 符合给定的规则 d 时输出 1，否则输出 0。而 $X_n=\{x_0, x_1, x_2,\cdots, x_n\}$, n∈N 表示某个观测序列，若存在 D_n，满足

$$\prod_{i=0}^{n} g(x_i, d_i) \equiv 1$$

则称 X_n 符合构词特征 D_n。此时目标观测集由规则序列生成，即

$$X_t \xrightarrow{g(x,d)} D_n$$

本文对构词特征的提取方式是，先将训练语料中的所有词汇合并为参考词典，再分别求出词典中任一两个词汇之间的最长公共字序列（longest common subsequence，LCS），并筛选出 LCS 足够大的词汇组合作为依据，利用该序列集生成有穷状态机（finite state machine，FSM）。该状态机将记录词典中出现的公共构词规则，利用这些规则，生成基于条件随机场的匹配特征函数，在识别中提供相应的匹配依据。

构词特征是基于构词规则的特征，而构词规则基于词典内容进行提取。利用这一特点，在训练过程中可以自定义词典内容，让模型抽取相应的构词特征，使得模型不局限在同类语料的识别中。

3 实验过程与结果分析

3.1 实验语料

既然是语言处理，自然需要相应的语料库。本文准备了三类语料库进行对比实验，这三类语料库涉及不同领域，其行文特点均有明显的差异。这些语料的共同点是均为汉语语料，其文本表达的架构词汇具备相同或相似的词汇特征，而包含的专有名词往往不同。

1)Bakeoff2005（SIGHAN）语料 Bakeoff 是 SIGHAN 所主办的国际中文语言处理竞赛，而主办方将其第二届(2005）比赛中的数据公开，用于支持各机构进行中文语言处理的研究。该语料也是目前被最广泛运用到的分词语料，其中，来自北京大学的数据为简体中文语料，其内容主要来源于人民日报的相关报道与资料，具有内容通俗易懂、表达大众化等特点。本文选取的正是其中来自北京大学的语料作为研究语料。

2)NLPIR 微博语料 该语料出处为 NLPIR。NLPIR 全称为自然语言处理与信息检索共享平台(http://www.nlpir.org/)。NLPIR 搜集了新浪微博上用户发表的内容作为语料内容。而该语料的特点是行文随意、情感丰富，其文本内容常常呈不完整的表达形式，属于语言特征比较不明显的汉语语料。

3)法律文书语料 本文搜集了人民法院公开发布的各类文书内容作为第三类研究语料。这些包括判决书、裁定书及执行通知书等人民法院发布的法律文件内容，其具有行文规范工整、表达逻辑性强、使用术语多等特点。

本研究中，抽取 Bakeoff2005 中的少许内容（共计 400 行）作为训练语料，而剩下语料内容及其他类型语料作为参考语料，判断相应特征对识别结果的影响。

3.2 BMES 标记集

本文将使用常用的 BMES 标记集对语料进行标记，不使用任何附加标记，尝试从最基本的标记中寻找词汇的特征。其中，标记 S 表示单字词或单标点符号，标记 B 表示多字词或多字符符号的首字符，M 表示中间字符，E 表示尾字符。在本文中，这些标记还有其他隐含的意义：B 表示当前词的开始标记，即先行词组合完毕；E 则表示当前词组合完毕，应以当前字作为词尾。

3.3 特征集符号

根据 2.2 节描述的几个词汇特征，本文通过考察与分析，决定在实验中将主要考察以下特征对识别结果的影响。特征集符号及描述如表 1 所示。

表 1 特征集符号及描述

Table 1 Descriptions of feature sets

符号	描述
R1	当前字的特征
R5	当前字以及前后 2 字（共 5 字）的特征集合
Pre	前缀的特征
Post	后缀的特征
Ant	先行词的特征
Sub	后继词的特征
MS	词开始字的特征
ME	词结束字的特征
D1	任意两词之间的 LCS ≥ 1 形成的构词特征
D2	任意两词之间的 LCS ≥ 2 形成的构词特征

3.4 评测方法

本文的评测基准为分词结果，并不对每个字的标签进行逐字评测。类似于命名实体识别的判别，本文使用自己的方式将标注结果转换成识别结果，并对识别结果进行判别，计算其准确率（Precision，P）、召回率（Recall，R）。本文在特征集的评测过程中主要考察模型识别的准确率，也因此不计算其 F 值（F-Measure）的大小。其中：

$$P = \frac{\text{准确识别出的词数}}{\text{识别出的总词数}}$$

$$R = \frac{\text{准确识别出的词数}}{\text{样本包含的总词数}}$$

3.5 模型效果对比实验

本文模型采用 CRF 原理构建，使用 L-BFGS 算法进行参数优化。在研究实验中，训练集、测试集内容均保持不变，只对模型的特征函数进行调整。其中，本文尝试了该模型（model）与开源工具 CRF++0.58（CRFPP）的识别效果进行了对比（表 2）。

从表 2 中对比结果来看，除去优化阈值、正则化参数等训练场景不同而造成的差异外，model 的识别效果基本与 CRFPP 的识别效果相似，随机特征（R）的增多对识别效果反而有负影响。由于训练集语料取自 BAKEOFF2005，所以其识别准确率、召回率均高于其他语料库的现象是合理的。

表 2 模型效果对比

Table 2 Model performances

模型	特征	Bakeoff		NLPIR		法律文书	
		P	R	P	R	P	R
CRFPP	R1	50.89	68.17	49.12	56.12	42.10	47.13
	R5	52.72	73.43	49.04	56.18	43.92	50.57
	R9	51.20	73.78	48.68	57.01	41.50	51.15
model	R1	50.37	67.33	48.52	54.46	40.66	46.31
	R5	50.06	71.91	47.81	54.56	41.41	49.84
	R9	48.52	72.63	45.68	54.17	39.39	50.59

从表中还可以看出，相比于成熟开源工具 CRF++，本文构建的模型在相同特征条件下略逊一筹，准确率差异最大达

到了 3%。本文之所以自行构建模型，是因为 CRF++无法满足本文想要自定义复杂特征函数的需求，且 CRF++在不引入额外标记的条件下，难以做到精确提取词汇的相关特征。

3.6　特征集综合对比实验

本文经过一系列实验，分别考察了不同类型特征的组合对模型的性能影响，每类选取平均准确率较高的前两种（共六种）组合进行综合对比，选取结果如表 3 所示。

表 3 选取的特征组合及识别准确率

Table 3　Precisions of selected feature sets

组合	类型	平均准确率
R1 + Pre + Post	词缀特征	0.48363
R5 + Pre + Post		0.48113
R1 + Ant + Sub	邻接特征	0.48150
R5 + Sub		0.47853
R5 + MS	边界特征	0.47017
R5 + ME		0.46973

根据表 3 的内容，本文对最终实验特征集进行组合，并进行对比实验比较其识别效果。特征集的选取结果如表 4 所示。

实验结果如图 8 表示。Set4、Set6 的识别效果要略优于其他特征集组合，其中 Set6 比 Set4 多了四种随机特征，但识别效果几乎没有变化。根据组合中的不同搭配可以看出，随机特征（R）只在有限范围内有利于识别效果；邻接特征中，先行词特征（Ant）在组合中反而会降低识别的准确率；结束字特征（ME）明显优于开始字特征（MS）。

表 4 特征集选取组合

Table 4　New sets of feature sets

特征集	组合
Set1	R1 + Pre + Post + Ant + Sub + MS
Set2	R1 + Pre + Post + Ant + Sub + ME
Set3	R1 + Pre + Post + Sub + MS
Set4	R1 + Pre + Post + Sub + ME
Set5	R5 + Pre + Post + Sub + MS
Set6	R5 + Pre + Post + Sub + ME

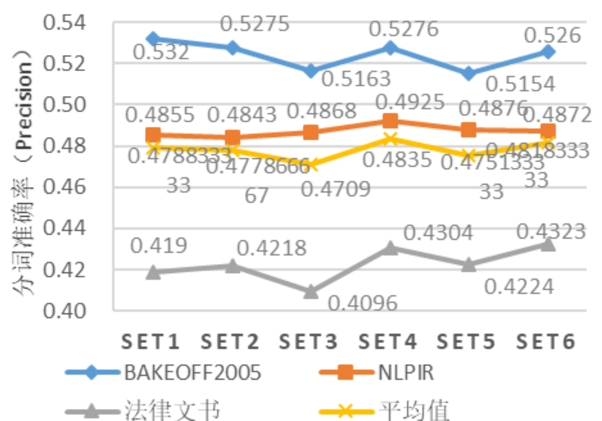


图 8　特征集组合对比结果

Fig. 8　Performances of new sets

3.7　构词特征对比实验

上一节中本文讨论了各种词汇相关的特征对识别的影响，得到了结果最优组合 Set4，即包含了随机特征（R1）、词前缀特征（Pre）、词后缀特征（Post）、后继词特征（Sub）、结束字特征 ME 的集合为效果最优的特征集合，通过这些特征，条件随机场能够很好地从训练集中提取相应词汇信息。

在本节中，将结合构词特征（D1、D2）分析更优的组合，并且在最后的实验中，增加训练集规模，对比不同组合对训练集的敏感程度。实验结果如图 9、10 所示。

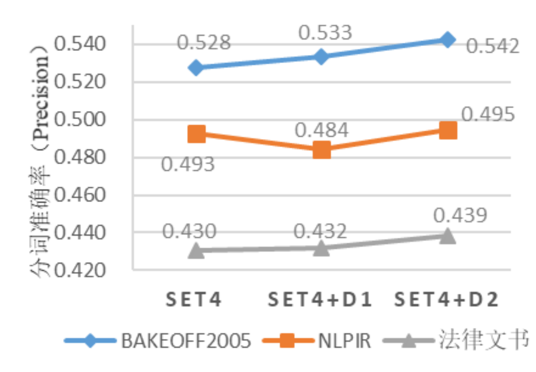


图 9　构词特征对比结果 1

Fig. 9　Performances of dictionary features 1

图 9 显示了 Set4 与构词特征组合后进行识别的准确率，结果表明，在特征集 Set4 添加 D2 特征能略微提高识别效果，而 D1 特征在 NLPIR 测试集中反而使得准确率降低，说明了构词特征确能在识别中提高识别效果，与其他特征可以共存。

图 10 则显示了增加训练集规模后，用相同特征集进行识别的结果。其中，增加的训练集与原训练集属于同类语料，即 Bakeoff2005 中的语料内容，增加后共计 1 001 行训练语料（原 400 行，增 601 行）。图 10 中，阴影填充的图例表示不同测试语料库的识别效果，而纯色图例表示增加语料库后准确率的增加量。可以看出，增加语料库后，使用三种特征组合集的识别效果均有提高，但组合 Set4 ＋ D2 有着更明显的性能提升。

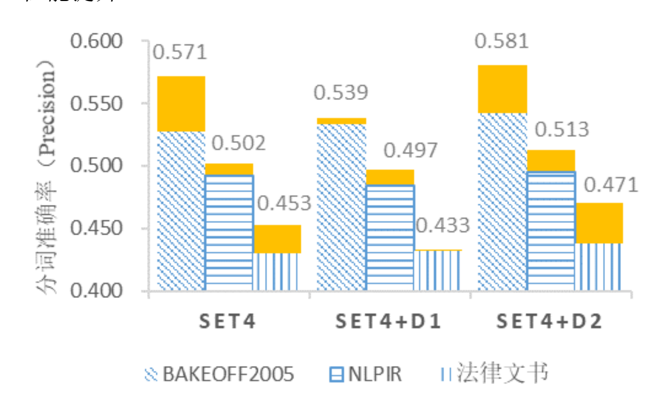


图 10　构词特征对比结果 2

Fig. 10　Performances of dictionary features 2

3.8　最终模型效果对比实验

在 3.5 节的模型对比实验中，本文构建的模型在 R5 特征下的性能略逊于开源工具 CRF++0.58 的性能。而通过对比与调整，本文找到了实验中结果相对最优的 Set4 ＋ D2（简称 S4D2）特征集，本小节亦将扩充后的训练语料应用于 CRF++0.58 中，再次进行模型对比实验（表 5）。

表 5 最终模型效果对比

Table 5　Performance comparisons of final feature sets

模型	特征	Bakeoff		NLPIR		法律文书	
		P	R	P	R	P	R
CRFPP	R5	52.72	73.43	49.04	56.18	43.92	50.57
	R5*	55.32	75.79	51.20	57.45	46.00	53.79
Model	S4D2	54.24	72.67	49.49	56.67	43.85	50.04
	S4D2*	58.06	75.29	51.27	57.21	47.05	52.94

表 5 的特征一列的表述中，加“*”号表示使用扩充后的训练语料（1001 行）进行实验，不加“*”号则表示实验使用扩充前的训练语料（400 行）。其中 CRFPP 表示使用开源工具 CRF++0.58，而 model 表示使用本文构建的条件随机场模型。

表 5 中显示了附加模型对比实验结果，其准确率对比也在图 11 与 12 分别给出。其中，本文选取了两模型在实验中性能最好的特征集进行对比（CRFPP 为 R5 特征集，而 MODEL 为 Set4+D2）。通过对比可以看出，经过对特征的选取，与 3.5 小节中的初步性能对比不同，model 已经达到了与 CRFPP 相同甚至更优的效果，扩充训练集后，model 在准确率上体现出了更好的变化。

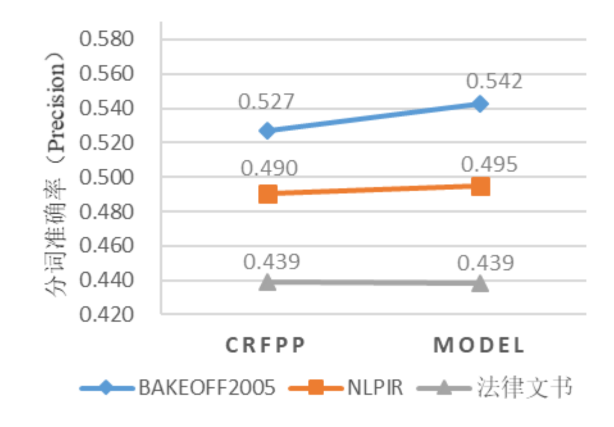


图 11 最优识别结果对比

Fig. 11 Best performance of two models

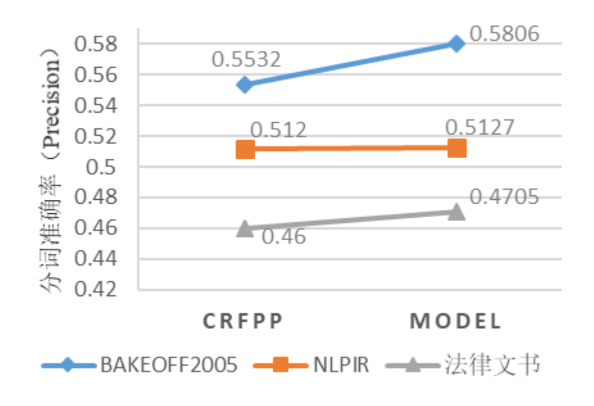


图 12 最优识别结果对比

Fig. 12 Best performance of two models

4 结束语

在命名实体的识别中，分词是一项重要的前置过程。本文研究了在基于简单标记（BMES 标记集）且不引入任何附加标记的条件下，如何提高词汇的识别准确率，提高算法可移植性。

通过实验，本文对比并组合了一些能够提高词汇识别效果的词汇特征，使用这些特征函数比一般特征有更好的效果。更重要的是，训练集中不需要提供任何附加标记，这也是解决语言处理算法可移植性问题中的一个重要方向。同时，本实验还说明了在条件随机场模型的应用中，灵活且符合情景的特征比随机而固定的特征更能保证模型的运行性能。本文的研究还说明了通过特定特征函数的选取，确实能够提高 CRF 在文本识别中的性能，提供了提高识别算法移植性的思路。

在今后的研究工作中，笔者将继续寻找更优的词汇特征，也会致力于模型优化算法的研究，寻找出更优而更有效率的识别特征集。

参考文献：

[1] 徐梓豪. 基于统计模型的中文命名实体识别方法研究及应用 [D]. 北京：北京化工大学，2017. (Xu Zihao. Statistical model based Chinese named entity recognition methods and its application to medical records [D]. Beijing: Beijing University of Chemical Technology, 2017)

[2] 万静，涂喆，冯晓. 基于条件随机场的医药领域症状信息抽取 [J]. 北京化工大学学报：自然科学版, 2016, 43 (1): 98-103. (Wan Jing, Tu Zhe, Feng Xiao. Information extraction of symptoms in the medical field based on CRF [J]. Journal of Beijing University of Chemical Technology: Natural Science Edition, 2016, 43 (1): 98-103.)

[3] 许旭阳，李弼程，张先飞，等. 基于条件随机场与自定义规则的时间表达式识别 [J]. 情报学报, 2011, 10 (10): 1065-1071. (Xu Xuyang, Li Bicheng, Zhang Xianfei, *et al.* Recognition of time expressions based on conditional random fields and rules [J]. Journal of The China Society for Scientific and Technical Information, 2011, 10 (10): 1065-1071.)

[4] 杨雁峰. 基于条件随机场和电商领域特征的中文产品描述信息自动分词研究 [D]. 上海：华东师范大学, 2015. (Yang Yanfeng. Product information words recognition based on conditional random field in electronic commerce [D]. Shanghai: East China Normal University, 2015)

[5] Lafferty J D, Mccallum A, Pereira F C N. Conditional random fields: probabilistic models for segmenting and labeling sequence data [C]// Proc of the 18th International Conference on Machine Learning. San Francisco: Morgan Kaufmann, 2001: 282-289.

[6] 汪光亚，曹国，尚岩峰. 基于条件随机场的多时相遥感影像分类 [J]. 计算机应用研究, 2018, 35 (9): 2834-2837. (Wang Guangya, Cao Guo, Shang Yanfeng. New multitemporal remote sensing image classification method based on conditional random fields [J]. Application Research of Computers, 2018, 35 (9): 2834-2837.)

[7] Mandal A, Ghosh K, Pal A, *et al.* Automatic catchphrase identification from legal court case documents [C]// Proc of ACM on Conference on Information and Knowledge Management. New York: ACM Press, 2017: 2187-2190.

[8] Sristy N B, Krishna N S, Krishna B S, *et al.* Language identification in mixed script [C]// Proc of the 9th Annual Meeting of the Forum for Information Retrieval Evaluation. New York: ACM Press, 2017: 14-20.

[9] Ramezani M, Khodadadi A, Rabiee H R. Community detection using diffusion information [J]. ACM Trans on Knowledge Discovery from Data, 2018, 12 (2): 1-22.

[10] 张祝玉，任飞亮，朱靖波. 基于条件随机场的中文命名实体识别特征比较研究 [C]// 第四届全国信息检索与内容安全学术会议论文集 (上) . 2008: 111-117. (Zhang Zhuyu, Ren Feiliang, Zhu Jingbo. A comparative study of features on CRF-based Chinese named entity recognition [C]// Proc of the 4th National Conference on Information Retrieval and Content Security (Part I) . 2008: 111-117.)

[11] 向晓雯. 基于条件随机场的中文命名实体识别 [D]. 厦门：厦门大学, 2006. (Xiang Xiaowen. Chinese Named entity recognition based on conditional random fields [D]. Xiamen: Xiamen University, 2006)

[12] 袁文宜. 依存语法概述 [J]. 科技情报开发与经济, 2010, 20 (18): 152-154. (Yuan Wenyi. A summary of dependency grammar [J]. Sci-Tech Information Development & Economy, 2010, 20 (18):

152-154.)

[13] 杨进才, 涂馨丹, 沈显君, 等. 基于依存关系规则的汉语复句关系词自动识别 [J]. 计算机应用研究, 2018, 35 (6): 1756-1760. (Yang Jingcai, Tu Xindan, Shen Xianjun, *et al.* Research on automatic recognition of relational word in Chinese compound sentences based on dependency relation rule [J]. Application Research of Computers, 2018, 35 (6): 1756-1760.)

[14] 邱立坤, 金澎, 王厚峰. 基于依存语法构建多视图汉语树库 [J]. 中文信息学报, 2015, 29 (3): 9-15. (Qiu Likun, Jin Peng, Wang Houfeng. A multi-view Chinese treebank based on dependency grammar [J]. Journal of Chinese Information Processing, 2015, 29 (3): 9-15.)

[15] 霍娟娟. 依存语法和配价语法在真词纠错中的研究与应用 [D]. 合肥: 中国科学技术大学, 2015. (Huo Juanjuan. Research and application of the dependency grammar and valence grammar in the real-word errors correction [D]. Hefei: University of Science and Technology of China, 2015)

[16] 孟禹光, 周俏丽, 张桂平, 等. 引入词性标记的基于语境相似度的词义消歧 [J]. 中文信息学报, 2018, 32 (8): 9-18. (Meng Yuguang, Zhou Qiaoli, Zhang Guiping, *et al.* Word sense disambiguation based on context similarity with POS tagging [J]. Journal of Chinese Information Processing, 2018, 32 (8): 9-18.)

[17] 吕凡. 基于生成对抗网络的图像自动文本标注方法研究 [D]. 苏州: 苏州科技大学, 2018. (Lyu Fan. Image captioning based on generative adversarial network [D]. Suzhou: Suzhou University of Science and Technology, 2018)

[18] 蒋婷. 学科领域本体学习及学术资源语义标注研究 [D]. 南京: 南京大学, 2017. (Jiang Ting. Discipline ontology learning and Semantic annotation for scientific resources [D]. Nanjing: Nanjing University, 2018.)